AD_____

Award Number: W81XWH-11-1-0549

TITLE: Enhancing the Breadth and Efficacy of Therapeutic Vaccines for Breast Cancer

PRINCIPAL INVESTIGATOR: Paul T. Spellman, Ph.D.

CONTRACTING ORGANIZATION: Oregon Health & Science University
Portland, OR 97239-3098

REPORT DATE: October 2013

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2013	Annual	25 Sep 2012 – 24 Sep 2013

4. TITLE AND SUBTITLE
Enhancing the Breadth and Efficacy of Therapeutic Vaccines for Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0549

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Paul Spellman, PhD; Dmitri Rozanov, PhD; Kami Chiotti, MS

E-Mail: spellmap@ohsu.edu, rozanov@ohsu.edu, chiotti@ohsu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Oregon Health & Science University
3181 SW Sam Jackson Park Road, Portland, OR 97239-3098

8. PERFORMING ORGANIZATION REPORT

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The focus of the Spellman/Gray work group over the past year has been upon the generation of materials, tools, and data for the purpose of aiding and supporting the research and findings of the entire multi-team collaboration endeavoring to identify antigenic targets for breast cancer-infiltrating T cells. During this time, MHC-I-bound epitopes for several of our breast cancer cell lines were ranked and evaluated to determine cell surface-presenting, MHC I-loaded epitopes from genes with preferential or altered expression in breast cancer cell lines. An analytical pipeline was also developed to perform *in silico* epitope prediction using RNAseq data as input; the data from which contributes to the understanding of expressed tumor-associated protein-encoding transcripts. Additional RNAseq analysis was carried out to determine HLA-A2 typing of our in-house breast cancer cell lines. Finally, a small molecule drug compound plates have been generated and are ready to determine enrolled patients' T cell IC50 for each compound.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	13	

Table of Contents

INTRODUCTION:
The OHSU Spellman/Gray work group is one of three collaborators funded by this Department of Defense Breast Cancer Multi-Team Award; the other two being comprised of the Lee work group from City of Hope (formerly of Stanford Medicine Cancer Institute) and the Slansky/Kappler work group from University of Colorado Denver/National Jewish Health. The major objective of this endeavor is to develop novel strategies aimed at the enhancement of the protective effects of anti-tumor T cells *in vivo* in a patient-specific manner based on the hypothesis that partially protective anti-tumor T cells exist within TDLNs in most breast cancer patients. This will be accomplished by identifying the antigens anti-tumor T cells target in different breast cancer subtypes, potentially including antigens preferentially expressed by breast cancer stem cells. We will identify both MHC-I- and MHC-II-restricted antigens driving both CD8 and CD4 anti-tumor T cells *in vivo*, as CD4 T cells are needed to optimally sustain vaccine-elicited CD8 T cells *in vivo* [1]. Identified antigens will be categorized as to breast cancer subtype-specificity or shared status amongst subtypes, with the intention a patient could be matched with an optimal set of vaccine antigens for her tumor. Another novel aspect of this project is the identification of altered peptides (mimotopes) that may more efficiently activate anti-tumor T cells than the natural tumor epitopes. A final objective is to identify small molecule anti-cancer agents that synergize with cytotoxic T lymphocytes (CTLs) to enhance immune-mediated killing. Collectively, this undertaking will produce a set of immunologically validated antigens and mimotopes for major breast cancer subtypes, and a set of agents that cooperate with immune killing. These can be used in combinations in a patient-specific manner to maximize clinical benefit while minimizing toxicity. The tools we develop will enhance the breadth and efficacy of existing and future approaches for immune therapy of breast cancer. We discuss here the Spellman/Gray group's specific efforts toward realizing the goals of this collaboration.

BODY:
Generation and initial analysis of T cell clones [Task 5]
As reported last year, the Spellman/Gray lab is contributing to the progress of this task through identification of MHC-I-restricted epitopes eluted from breast carcinoma cell lines utilizing a combination of immunocytochemistry, immunoprecipitation and mass spectrometry. Our *in vitro* model of breast cancer is a diverse collection of 70 breast cancer cell lines, which are the focus of intensive molecular and phenotypic characterization. We used these breast carcinoma cell lines to determine the sequence and the level of MHC-I-bound epitopes expressed on the cell surface, constructing a comprehensive panel of confirmed epitope sequences.

In brief review of our procedure, we first identified MHC-I-positive breast carcinoma cells (MDA-MB-231, SUM159PT, CAMA-1, MCF7) by staining with MHC-I pan-specific and A2 subtype-specific antibodies. Nonspecific Ms-IgG staining was used as a negative control. Next, we developed a very efficient procedure (as detailed in the 2012 annual report) to immunoprecipitate MHC-I molecules followed by elution of MHC-I-bound epitopes with trifluoroacetic acid (TFA), allowing us to identify MHC-I-restricted epitopes expressed on the surface of different breast carcinoma cells. The sequences of the peptides bound to MHC-I were acquired following analysis by mass spectrometry

The total number of eluted peptides from the cell surface and the corresponding number of proteins associated with those peptides is equal to 3366 and 3078, respectively. This number does not correspond to the number of unique peptides and proteins (Table 1) because there are shared MHC I-presented peptides and proteins among different breast carcinoma cell lines. After removing duplicates, the numbers of unique epitopes and corresponding proteins is 2821 and 1940, respectively.

To find breast cancer specific MHC I-loaded epitopes that could have the ability to activate T cell response, we used gene expression profiling to determine the MHC I-presented genes with alterations or elevated expression levels in breast tumors compared to normal cells. First, we determined genes whose expression is altered in invasive breast cancers by copy number amplification, homozygous deletion, mRNA upregulation or downregulation, and mutation using the cBioPortal for Cancer Genomics that

	Cell line	Subtype	№ peptides	FDR, %	№ proteins	FDR, %
1	SUM159PT	Claudin-law	**439**	13	**385**	13
2	MDA-MB-231	Claudin-law	**9**	10	**9**	9
2	MDA-MB-231	Claudin-law	**49**	15	**46**	15
2	MDA-MB-231	Claudin-law	**10**	6	**10**	20
3	HCC1395	Claudin-law	**83**	9	**81**	10
4	BT549	Claudin-law	**22**	1	**22**	20
5	HCC70	Basal	**271**	9	**251**	8
6	HCC1187	Basal	**688**	6	**607**	9
7	HCC1569	Basal	**200**	6	**189**	9
8	MCF12A	Basal	**87**	1	**83**	4
9	CAL-120	Basal	**4**	1	**4**	11
10	HCC1500	Basal	**33**	8	**32**	9
11	MDA-MB-468	Basal	**274**	6	**256**	7
12	HCC1806	Basal	**299**	6	**273**	9
13	LY2	Luminal	**241**	5	**226**	11
14	MCF7	Luminal	**222**	6	**203**	9
15	CAMA-1	Luminal	**118**	1	**104**	4
16	T47D HER2+	Luminal	**75**	1	**71**	9
17	HCC1419	Luminal	**17**	2	**17**	10
18	HCC1428	Luminal	**22**	1	**21**	7
19	SUM185PE	Luminal	**88**	2	**86**	2
20	UACC812	Luminal	**107**	2	**94**	3
	Total		**3358**		**3070**	
	Unique		**2813**		**1939**	

Table 3. Number of eluted MHC I- restricted peptides and corresponding proteins in breast carcinoma cells. (FDR=false discovery rate).

contains large-scale cancer genomics data sets. We arranged all identified genes in accordance with the frequency of alterations in breast cancer samples. For further analysis we selected genes that have alterations in at least 20% of breast cancers.

We then used gene expression data for 708 breast tumors and 329 normal tissues from The Cancer Genome Atlas (TCGA) [2], the European Bioinformatics Institute (EBI) [3], and the Gene Expression Omnibus (GEO) [4] to identify among the MHC I-presented genes those genes having preferential expression in breast cancer samples over normal samples. Alignment and expression values were generated using the Myrna software package [5]. We averaged expression amongst all tumor and normal samples for each gene and ranked the genes by level of differential expression in tumor and normal samples. In this analysis, we selected genes with at least 4 times higher expression in cancers than in normal tissues.

Using the same data set, we evaluated differential expression across all normal and tumor samples by calculating the Median Split Silhouette (MSS) of each gene. MSS is a clustering algorithm measuring the average heterogeneity of possible clusters and determines whether the expression profile of a gene, across all normal and tumor samples, is best described by one or more clusters [6]. The advantage of MSS comes from its ability to identify biologically meaningful clusters where cluster size may be small. For our purposes, we limited the maximum number of potential clusters to three (*kmax*=3). This *kmax* was chosen in an effort to capture separation of gene expression between normal and tumor tissues as well as any bimodal expression amongst the tumor samples alone [7]. Of the nearly 2000 genes identified following immunoprecipitation and elution of their associated epitopes, MSS predicted 494 of

the genes to cluster into two or three expression groups. The remaining genes were either predicted to display only one expression cluster (i.e., no potential of discerning tumor and normal expression profiles) or there was no expression information collected by Myrna (i.e., no reads aligned to the gene). Using this clustering, we selected 26 genes demonstrating preferential expression in breast tumors.

We attempted to select breast cancer specific candidate genes using RNAseq data from 62 breast carcinoma cell lines and 6 non-transformed cell lines. We averaged expression data for each gene across all breast carcinoma cell lines and non-transformed cells, and for further analysis, we selected genes with 4 times higher expression in transformed over non-transformed cells.

As an additional approach to identify immunogenic genes, we looked for genes frequently identified by our MHC I immunoprecipitation and elution approach among different cell lines. We arranged all MHC I-presented genes based on the number of times each gene was identified among cell lines of a particular subtype and among all cell lines. We selected genes that have been identified at least 10 times in all analyzed cell lines or at least 5 times in subtype-specific cell lines. In addition, because the HLA-A2 allele is frequently present in all ethnic groups [8], we limited our analysis to MHC I-presented genes identified in HLA-A2-positive breast carcinoma cells. The ability of the selected peptides to be loaded into the peptide binding groove of HLA-A2 molecules was confirmed by the high binding score calculated by an epitope prediction algorithm [9]. These activities allowed us to select 132 MHC I-loaded epitopes from genes exhibiting either preferential or altered expression in breast cancers and breast carcinoma cells and are frequently presented on the surface of the analyzed cells.

Additionally, optimization of conditions for amplification of the T cell receptor (TCR) gene using total RNA sample from breast cancer patients was carried out. We employed a template-switching approach and step-out PCR to amplify TCR cDNA 5'-end of the unknown sequence [10]. We were able to amplify the variable region of TCR-alpha but not that of TCR-beta. We have decided to use the published protocol for TCR cDNA amplification from a single cell [11].

RNAseq analysis of tumor cells [Task 7]

RNAseq analysis to identify breast cancer-specific aberrant transcripts. RNAseq datasets are being used to conduct a systematic computational analysis to identify aberrant transcripts resulting in breast cancer antigens. The Spellman/Gray computational group has developed an epitope prediction pipeline utilizing approximately 1000 breast cancer and normal tissue RNAseq samples available through TCGA, EBI, and GEO. Over one-third of the RNAseq samples originated from normal adult tissues, predominantly made up of breast, lung, liver, brain, heart, kidney, and B-cells. A variety of other tissues are also represented, albeit in smaller sample numbers, to include bowel, skeletal muscle, lymph node, and ovary, amongst others. The entirety of the tumor dataset was obtained from the TCGA Data Portal. Of the better than 700 tumor samples, TCGA categorized approximately 460 samples into basal, Her2, and luminal subtypes using the PAM-50 subtype prediction method [12]. Only sequences generated on the Illumina Genome Analyzer II and Genome Analyzer IIx [13] platforms were included in the study to maintain as much uniformity as possible between datasets generated at different locations. As many of the sequences were single-end reads and read lengths varied from 50-150bp, all paired-end sequences were converted to single-end, and read lengths were trimmed as necessary to 50bp prior to being submitted in the form of FASTQ files to the analytical pipeline depicted in Figure 1.

Mining of the RNAseq dataset was initiated through implementation of the Bowtie/Tophat/Cufflinks [14]–[16] packages (collectively referred to as the Tuxedo suite) to carry out sequence assembly and alignment to the human genome (hg19), prediction of novel isoforms, and quantitation of transcript structure. Using the Cuffmerge [16] feature of Cufflinks, the entire set of assemblies were merged such that identical transcripts across all samples were accounted for by a single identifier and its associated gene expression values.

Novel isoforms of a transcript can indicate alternative splicing events not yet characterized by the reference genome as well as aberrant structural variations due to mutation, both of which can result in

neoantigens. Due to very low representation of the novel isoforms in some samples, it is likely the Tuxedo suite may not have detected, assembled, and subsequently determined the expression level for the new isoform in every sample. In order to force Tuxedo to look for and calculate the expression values of all isoforms in each sample, the subset of transcripts predicted to be novel assemblies were extracted from the Cuffmerge output and used to construct a new transcriptome index. The entire RNAseq dataset was rerun through the Tuxedo suite using this new index as the reference sequence. From here on, the collections of native and novel transcripts are kept separate from each other but run in parallel through the remainder of the pipeline.

For calculation of gene expression levels, we used the binary logarithm of the FPKM (fragments per kilobase of transcript per million mapped reads) values as calculated by Cufflinks. The FPKM values underwent full-quantile normalization utilizing the *betweenLaneNormalization* function of the EDASeq R/Bioconductor package [17]. This function accounts for distribution differences by matching the quantiles of the count distributions between samples as described in [17] and [18]. Differential expression of the genes was then determined utilizing the MSS clustering discussed previously (Task 5). Filtering steps were then taken to winnow the dataset to only those genes found within the high-expression cluster representing (1) at least an eight-fold expression differential between high- and low-expression clusters, (2) a large tumor population (>95% tumor within the cluster) **and** (3) a significant portion of the total tumor population (>10% of all tumor samples in the dataset). This resulted in narrowing the native transcript candidates from ~79K to ~175 and the novel isoform candidates from ~116K to ~185. The reasoning behind this filtering scheme is as follows:

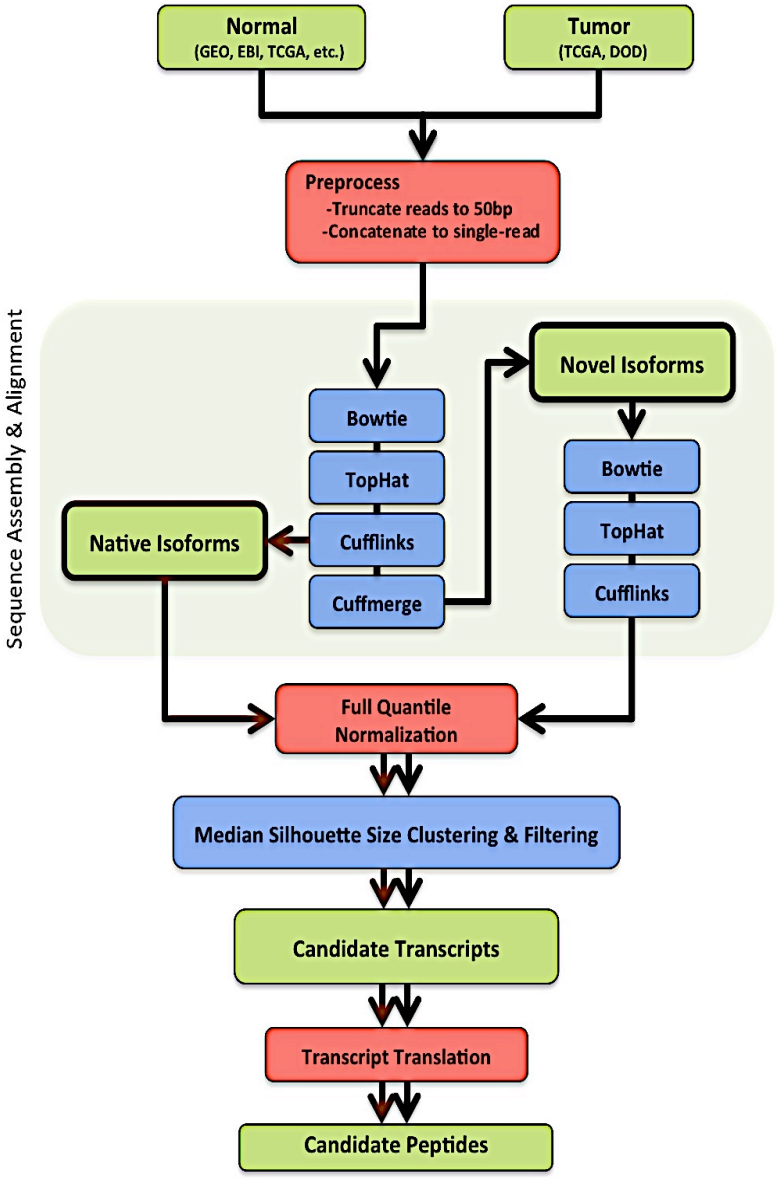

Figure 1. Pipeline for analysis of RNAseq data to identify native and neoantigen sequences.

1. The epitope must be expressed at a significantly higher level in tumor tissue than in normal tissue in order to be immunologically targetable.
2. The epitope must be specific to breast tumor to avoid inadvertently targeting and damaging normal tissue.
3. The epitope should be targetable in a significant portion of the breast cancer population

With evidence of approximately 185 differentially expressed novel transcripts identified amongst the

dataset, it was necessary to designate those which held the most potential for translation into unique peptide constructs. As it was necessary to accomplish this task manually, we initially removed any isoforms indicating start and stop sites at the 5' and 3' ends, respectively, which were identical to their nearest known reference transcript. This step narrowed the total number of transcripts needing manual validation to 51, retaining those with the most variations as compared to known transcripts. We will return to the isoforms removed during this step in the future to determine whether internal variations are present with the potential of translation to a novel epitope.

At this point, the coding sequence of each unique transcript as predicted by the Tuxedo suite was translated to its corresponding peptide sequence using the TranSeq tool [19], [20] in all three frames. The most likely reading frame was selected via alignment to the human reference genome (hg19) using the UCSC-BLAT web tool [21]. The novel transcript nucleotide sequences were also aligned to hg19 utilizing UCSC-BLAT to visually confirm the accuracy of the nearest predicted reference transcript as determined by Tuxedo. An additional web tool, Clustal Omega [22], was then used in which the predicted nucleotide sequence was aligned to the nearest reference coding sequence. Similarly, the translated novel peptide sequence was aligned to the nearest reference peptide sequence. In those cases where the Tuxedo-predicted nearest reference did not produce the best alignment, it was replaced by the more appropriate sequence. Manual cross-comparison of the UCSC-BLAT and the two Clustal Omega alignments was carried out to reveal the most likely coding sequence of the predicted novel isoform. All isoform variations demonstrating the potential of producing an alternate start or stop translation site, an inclusion or exclusion of whole or partial exons, or a combination of exons unique amongst all known reference transcripts were documented.

To be relevant as an immunological target, the epitope must be expressed at a significantly higher level in tumor tissue than normal tissue. Those native transcripts *preferentially* expressed in tumors (high-expression cluster contains >95% tumors and represents >10% of the tumor population) and demonstrating the highest expression levels (greater than 8-fold difference from nearest neighboring cluster) include the genes CD44, CREB3L4, FIP1L1, KCN34, MAZ, P4HA3, PIGF, PUSL1, RBM17, BMPR1B, TMEM150C, OBP2B, and two transcripts each of NAT1 and STARD10. Tumor-*specific* transcripts (tumor population of high-expression cluster population is 100% and >4-fold difference from nearest neighboring cluster) include EN1, S100A7, SLITRK6, COL2A1, CST9, CST1, MMP11, IL20, RET, and FCRLB. These genes are of particular interest due to their reduced potential of vaccine cross-reactivity with normal tissue. Evaluation of tumors of known subtype also reveals evidence of differential expression amongst the subtypes. Her2 and luminal tumors are found to preferentially express AGR2, DEGS2, and TPD52 transcripts. Overexpression of these transcripts is found in 78-92% of the Her2 and 85-92% of the luminal tumors in the dataset compared to only 6-15% in basal. Two different NAT1 transcripts exhibit preferential expression in 74-81% of luminal tumors, but 24-26% and 2.3% of Her2 and basal tumors, respectively. Approximately 91% of the basal and 68% of the Her2 tumors express FOXM1 at significant levels as opposed to 16% of the luminal tumors. The above-noted EN1 transcript also shows higher expression amongst 67% of the basal samples, while only 6% and 0.7% of the Her2 and luminal samples are significantly expressed, respectively.

Novel tumor-*specific* isoforms with high-expression clusters at least 4-fold greater than the nearest neighboring cluster include those most closely related to a known transcript of CASP14, UNC5C, COL11A1, COL12A1, CST1, NCCRP1, or TPRG1. MMP11 and TPRG1 each have two novel isoforms for the same reference transcript meeting these criteria. The CST5 novel isoform is *preferentially* expressed in tumors with the high-expression cluster over 96% tumor. A third novel isoform of TPRG1 has a high-expression cluster tumor population of 99.5%. Finally, an SPDEF novel isoform exhibits differential expression amongst breast cancer subtypes where significantly overexpressing samples consist of 99.3% of all luminal samples and 100% of all Her2 samples in the dataset; however only 21.3% of the basal samples overexpress this isoform.

The majority of the genes discussed here have been identified in previous breast cancer studies, lending

support to the functionality of our pipeline and validity to our results. These results can be used to help guide future research for immunological targets, and the computational procedure can be used with enrolled patient RNAseq data to verify and quantitate novel isoform expression.

RNAseq analysis of breast carcinoma cell lines. To determine the type of MHC I alleles in the breast carcinoma cells, which we used in our MHC I elution studies, we completed RNAseq experiments for 18 cell lines. RNAseq analysis was based on paired end reads of at least 75 bp and at least 6 Gbp of mappable sequence. We used the recently published seq2HLA method [23] to map RNAseq reads against a reference database of HLA alleles. The HLA type I allele assignments and their associated p-values are listed in Table 2. Interestingly, when we looked at the RNAseq data we found that MHC I mRNA often does not have exons 1, 2, 5, 6 and 7 (Appendix I). It is known deletion of exons 6 and 7, which encode the cytoplasmic portion of MHC I molecules, drastically impairs proper MHC I trafficking through endosomal and lysosomal compartments and cytotoxic T lymphocyte (CTL) responses *in vivo* [24].

		Deleted exons		
Cell line		**HLA-A**	**HLA-B**	**HLA-C**
1	MDA-MB-231	6 and 7	1, 2, 6, and 7	6 and 7, partially
2	MDA-MB-468			
3	CAMA-1	6 and 7	2, partially	
4	BT549		6 and 7	6 and 7
5	HCC70			
6	HCC1395	5 and 6		7
7	HCC1419	6 and 7	2, partially, 6, and 7, partially	7
8	HCC1428	6 and 7	6 and 7	
9	HCC1500	6, partially		7, partially
10	HCC1569			7, partially
11	HCC1806		7, partially	6, partially and 7, partially

Table 2. Exon deletions in the HLA-A, B, and C genes in breast carcinoma cell lines.

Identify small molecule agents enhancing tumor cell apoptosis and CTL killing [Task 12]
As outlined in Aim 4 of the proposal, clinical efficacy of T cell-based therapies will be enhanced in combination with agents promoting tumor cell apoptosis. Support for this idea recently has been published showing chemotherapy can synergize with CTL-mediated killing [25]; however, chemotherapeutic agents can also inhibit T cell function. In order to identify drugs nontoxic to normal cells, we designed and ran cytotoxicity assays using three normal T cell clones from breast cancer patients and a collection of FDA-approved drugs consisting of 63 compounds during funded year one. All assays were done in triplicate at nine concentrations. Standardized compound plates have been created and are ready to determine IC50 for each compound against enrolled patient T cells. In addition, we have optimized medium composition (concentration of each: IL-2, IL-7, anti-CD3/anti-CD28 Macsi beads, human serum) to propagate T cells.

KEY RESEARCH ACCOMPLISHMENTS:

- Identified 132 MHC I-loaded epitopes, frequently presented on the cell surface, from genes with preferential or altered expression in breast cancers and breast cancer cell lines.
- Created an analytical pipeline for *in silico* prediction of breast cancer epitopes from RNAseq data, identifying approximately 175 native transcripts and approximately 50 novel splice variants specifically or preferentially expressed in breast cancer tissue.

- Performed HLA-A2 typing of in-house breast cancer cell lines from RNAseq data.
- Constructed 63-compound cytotoxic assay plates for pending screening of enrolled patient T cells.

REPORTABLE OUTCOMES:

- **NBCC/Artemis Project:** We have completed our portion of the Artemis Project®, which was launched by the National Breast Cancer Coalition (NBCC) in September 2010 as a strategic campaign to end breast cancer by the end of the decade. The ultimate goal of the Artemis Project® is to help open the door to personalized breast cancer immunotherapy and promote development of a preventative vaccination for breast cancer. Our proposed project sought to develop a robust portfolio of native and non-native antigens across the major breast cancer subtypes using strictly computational means.

CONCLUSION:

The focus of the Spellman/Gray work group over the past year has been upon the generation of materials, tools, and data for the purpose of aiding and supporting the research and findings of the entire multi-team collaboration endeavoring to identify antigenic targets for breast cancer-infiltrating T cells. We have identified a number of candidates in breast cancer tissues as well as breast cancer cell lines, utilizing a variety of analytical methods. The RNAseq analysis tool is proof of concept of *in silico* epitope discovery from RNAseq data. It aids in the definition of the protein-epitope relationship by enlarging the knowledge base of protein-encoding transcripts beyond the protein models existing in public databases and by restricting the analyses to only the expressed transcripts. The results produced by this pipeline along with the MHC-I-bound epitopes identified by mass spectrometry in breast cancer cell lines will be used to rank epitopes for further characterization and development as therapeutic targets.

REFERENCES:

[1] A. J. Zajac, K. Murali-Krishna, J. N. Blattman, and R. Ahmed, "Therapeutic vaccination against chronic viral infection: the importance of cooperation between CD4+ and CD8+ T cells," *Curr. Opin. Immunol.*, vol. 10, no. 4, pp. 444–449, Aug. 1998.

[2] "Home - The Cancer Genome Atlas - Cancer Genome - TCGA." [Online]. Available: http://cancergenome.nih.gov/. [Accessed: 26-Mar-2013].

[3] M. Kapushesky, I. Emam, E. Holloway, P. Kurnosov, A. Zorin, J. Malone, G. Rustici, E. Williams, H. Parkinson, and A. Brazma, "Gene Expression Atlas at the European Bioinformatics Institute," *Nucleic Acids Research*, vol. 38, no. Database, pp. D690–D698, Nov. 2009.

[4] R. Edgar, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, Jan. 2002.

[5] B. Langmead, K. D. Hansen, and J. T. Leek, "Cloud-scale RNA-sequencing differential expression analysis with Myrna," *Genome Biology*, vol. 11, no. 8, p. R83, 2010.

[6] K. S. Pollard and M. J. Van Der Laan, "A method to identify significant clusters in gene expression data," *UC Berkeley Division of Biostatistics Working Paper Series*, no. Working Paper 107, Apr. 2002.

[7] M. Bessarabova, E. Kirillov, W. Shi, A. Bugrim, Y. Nikolsky, and T. Nikolskaya, "Bimodal gene expression patterns in breast cancer," *BMC Genomics*, vol. 11 Suppl 1, p. S8, 2010.

[8] J. M. Ellis, V. Henson, R. Slack, J. Ng, R. J. Hartzman, and C. Katovich Hurley, "Frequencies of HLA-A2 alleles in five U.S. population groups: Predominance of A∗02011 and identification of HLA-A∗0231," *Human Immunology*, vol. 61, no. 3, pp. 334–340, Mar. 2000.

[9] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanović, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, no. 3–4, pp. 213–219, Nov. 1999.

[10] M. Matz, D. Shagin, E. Bogdanova, O. Britanova, S. Lukyanov, L. Diatchenko, and A. Chenchik, "Amplification of cDNA ends based on template-switching effect and step-out PCR.," *Nucleic Acids Res*, vol. 27, no. 6, pp. 1558–1560, Mar. 1999.

[11] T. Ozawa, H. Kishi, and A. Muraguchi, "Amplification and analysis of cDNA generated from a single cell by 5'-RACE: application to isolation of antibody heavy and light chain variable gene sequences from single B cells," *BioTechniques*, vol. 40, no. 4, pp. 469–470, 472, 474 passim, Apr. 2006.

[12] D. C. Koboldt, et al, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Sep. 2012.

[13] "Illumina," 2013. [Online]. Available: www.illumina.com.

[14] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.

[15] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, Mar. 2009.

[16] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, May 2010.

[17] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit, "GC-Content Normalization for RNA-Seq Data," *U.C. Berkeley Division of Biostatistics Working Paper Series*, Aug. 2011.

[18] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics*, vol. 11, no. 1, p. 94, 2010.

[19] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, Jun. 2000.

[20] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, Juri Paern, and R. Lopez, "A new bioinformatics analysis tools framework at EMBL-EBI," *Nucleic Acids Res.*, vol. 38, no. Web Server issue, pp. W695–699, Jul. 2010.

[21] W. J. Kent, "BLAT—The BLAST-Like Alignment Tool," *Genome Res.*, vol. 12, no. 4, pp. 656–664, Apr. 2002.

[22] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular Systems Biology*, vol. 7, Oct. 2011.

[23] S. Boegel, M. Löwer, M. Schäfer, T. Bukur, J. de Graaf, V. Boisguérin, O. Türeci, M. Diken, J. C. Castle, and U. Sahin, "HLA typing from RNA-Seq sequence reads," *Genome Med*, vol. 4, no. 12, p. 102, Dec. 2013.

[24] G. Basha, G. Lizée, A. T. Reinicke, R. P. Seipp, K. D. Omilusik, and W. A. Jefferies, "MHC Class I Endosomal and Lysosomal Trafficking Coincides with Exogenous Antigen Loading in Dendritic Cells," *PLoS ONE*, vol. 3, no. 9, p. e3247, Sep. 2008.

[25] R. Ramakrishnan, D. Assudani, S. Nagaraj, T. Hunter, H.-I. Cho, S. Antonia, S. Altiok, E. Celis, and D. I. Gabrilovich, "Chemotherapy enhances tumor cell susceptibility to CTL-mediated killing during cancer immunotherapy in mice," *J. Clin. Invest.*, vol. 120, no. 4, pp. 1111–1124, Apr. 2010.

APPENDICES:

Appendix I. HLA genotyping from RNA-seq data in breast carcinoma cell lines. HLA genotypes and associated p-values from seq2HLA algorithm are shown for each cell line. HLA-A2 phenotypes determined by ICC analysis are also shown below the names of cell lines.

Cell Line	HLA	HLA1	p-Value	HLA2	p-Value
MDAMB231	A	A*02	0	hoz("A*24")	0.0101329
A2+	B	B*41	9.27E-08	B*40	0.013585
	C	C*17	3.90E-14	C*02	0.01060021
MDAMB468	A	A*23	0.00201634	A*30	0.000300322
A2-	B	B*27	0.00465698	B*53	0.002051244
	C	C*02	0.00031031	C*04	0.0427023
CAMA1	A	A*02	0	A*32	0.01403118
A2-	B	B*40	0.00010729	B*15	0.05666284
	C	C*02	3.84E-05	C*03	0.007110044
BT549	A	A*01	0.00010431	A*02	0.001187262
A2+	B	B*15	1.11E-14	B*56	0.4242975
	C	C*07	0	C*03	0.009379311
HCC70	A	A*30	0	A*03	0.002774575
A2-	B	B*78	3.79E-09	B*15	3.64E-05
	C	C*16	2.21E-07	hoz("C*03")	0.0002355
HCC1395	A	A*29	0	hoz("A*31")	0.2863664
A2-	B	B*08	0.00025113	B*45	0.001237614
	C	C*07	3.20E-08	C*06	0.01455983
HCC1419	A	A*24	0.00067757	A*02	0.03597097
A2-	B	B*46	0.04772688	B*52	0.03344892
	C	C*03	0.00332044	C*01	0.00971956
HCC1428	A	A*01	0.00611302	A*02	0.01550328
A2-	B	B*07	5.09E-08	hoz("B*35")	0.8173494
	C	C*07	0	hoz("C*12")	0.0014498
HCC1500	A	A*68	7.57E-11	A*23	0.01166152
A2+	B	B*51	0.00010822	B*15	0.000209064
	C	C*02	0	hoz("C*04")	2.70E-05
HCC1569	A	A*30	0	A*68	0.003631942
A2-	B	B*58	1.04E-05	B*53	0.004338769
	C	C*04	0.00665832	C*15	0.01074122
HCC1806	A	A*68	2.54E-08	A*23	0.007613425
A2-	B	B*51	4.08E-05	B*15	0.000383347
	C	C*02	0	hoz("C*14")	1.96E-05
LY2	A	A*02	0	hoz("A*33")	0.7782346
A2+	B	B*44	1.71E-13	B*18	0.01163663
	C	C*05	0.00395624	hoz("C*06")	4.11E-05
MCF7	A	A*02	0	hoz("A*24")	1
A2+	B	B*44	0	hoz("B*35")	1
	C	C*05	0.00698762	hoz("C*04")	2.23E-05

T47D	A	A*33	3.87E-11	hoz("A*11")	0.0504919
A2-	B	B*14	0	hoz("B*51")	0.4009827
	C	C*08	0.03627429	hoz("C*12")	7.50E-06
UACC812	A	A*68	0	A*02	0.000337381
A2+	B	B*51	0.00098582	B*15	0.003599431
	C	C*08	0.00201796	C*12	0.06304602
HCC1187	A	A*31	2.82E-05	A*01	0.02122619
A2+	B	B*08	0.00088782	B*40	0.007880993
	C	C*07	0.00086718	C*03	0.02039814
SUM159PT	A	A*24	0.00163221	A*02	0.002906545
A2+	B	B*51	0.02056039	B*15	0.00892504
	C	C*15	1.93E-05	C*03	0.006169382
MCF12A	A	A*66	0.02536765	A*02	0.000909216
A2+	B	B*41	0.00362804	B*35	0.002059722
	C	C*17	0.01016061	C*07	0.01017766